Analyzing Retention from Participation in Co-curricular Activities

Roger Blake  
University of Massachusetts Boston  
Management Science and Information Systems  
100 Morrissey Blvd  
Boston, MA 02125-3393  
617-287-7692  
roger.blake@umb.edu

Paul Mangiameli  
College of Business Administration  
University of Rhode Island  
Kingston, RI 02881-0802  
401-884-4217  
mangia@uri.edu

ABSTRACT

The impact of student retention is widely recognized by colleges and universities as extending well beyond financial, and retention is an issue that researchers have increasingly been turning to. Recently many have been focused on analytic methods, especially predictive models based on data mining techniques. Generally, these models use data on demographics, academic performance, socio-economic backgrounds, standardized tests, or financial aid. Student engagement is known to influence retention, but also can be more difficult to measure and is used less often. When student engagement is used, it is generally represented by the extent to which student support services such as advising are utilized. We explore a new approach to analyzing the influence of student engagement on retention and demonstrate how data mining models for retention might be developed based on data from participation in a program of co-curricular activities named MAP. Our research uses cluster analysis to show there are three distinctly different patterns of participation, and that participants tend to be in one of three groups, which we termed “High achievers”, “Meets expectations”, and “Retention risks.” We find statistically significant differences between each of these patterns and how those individuals most likely to be at risk of attrition might be identified.

INTRODUCTION

Student retention has an enormous impact on colleges and universities that goes well beyond the financial. Retention and student attrition affect the reputation, rankings, culture, enrollment decisions, and the viability of educational institutions. Even with the recognition of how importance retention is and considerable efforts to address it, the percentage of students who enroll and complete their undergraduate degrees has changed little [9], and nationwide the percentage of students who complete a four year degree is only 35.3% for public four year
institutions and 40.7% for those that are private [1]. It is then not a surprise that researchers have increasingly turned their attention to this topic. The methodologies most used in these studies are surveys and data mining; of these, data mining has recently been gaining prevalence.

Prior research of predictive data mining models for retention has generally used a combination of data on academic performance, demographics, test scores, socio-economic background, and family education levels. While student engagement is known to influence retention, when it has been used in data mining models, it has generally been through measures of the utilization of student support services such as tutoring or academic advising. Our research explores a new approach to evaluating how student engagement influences by incorporating participation in a program of co-curricular activities.

This paper explores how participation in that program, named the Management Achievement Program, or MAP, might be used to categorize and predict student retention. The purpose of MAP is a program is to build professional skills and student engagement, and that end it has been successful. MAP is required for all students, and because of that, the opportunity to test whether MAP increases the sense of involvement or engagement is limited – there is no control group for such a test. However, as part of administering MAP an array of data regarding participation and events is gathered. This research investigates how that data might be utilized for data mining models to predict retention, and models potentially capable of making that determination very soon after enrollment when attrition rates are known to be highest and accurate predictions the most valuable.

BACKGROUND

Prior work
Data mining models used to predict retention from prior research are most often built with data for a combination of demographics, financial aid, family educational backgrounds, performance measures, standardized tests, and faculty experience. For example, Nandeshwar, Menzies, and Nelson [8] note there is much room for improving predictive models of retention and provide a comprehensive review of prior research. They hypothesize three factors as having significant influences on retention: financial aid, academic performance, and faculty tenure and experience, finding socio-economic and academic performance to be most significant. In another example, starting with demographic data such as age, performance measures such as SAT scores, chosen majors, work study history, financial aid data, and data on scholarships, Defen [3] focused on ensemble methods to evaluate a suite of data mining algorithms. In concluding which ensemble methods are better, Defen commented that the results from the models might have been improved if more data were available, particularly for social interaction. The participation data used for our research may be useful in this regard because a number of MAP events involve student clubs and volunteer activities, although we hold that investigation for further research.

There is considerable evidence that student engagement influences outcomes including critical thinking, cognitive development, self-esteem, satisfaction, and the factor that is most central to our research, retention [13]. When student engagement has been incorporated in data mining models of retention, engagement has often been represented by the extent to which student
support services have been utilized. For example, several models were compared using data on support service utilization and publicly available data on demographics, family education levels, course enrollments, and financial aid. That study ultimately concluded that utilizing student support services and successful completion of foundation courses were the most significant influences on retention [4].

Another example of measuring student engagement through the utilization of services, and a precedent for using cluster analysis to analyze its patterns as we do, is from a typology based on clustering data on utilizing student services such as academic advising, skill labs, and student organizations, combined with demographic data for 320,000 students [11]. With a much wider scope and population for analysis, the typology from their cluster analysis consisted of 15 clusters rather than the three found from ours. However, like our research, distinct clusters of participation could be identified, and like our research, the significance of these patterns for analyzing and improvement retention was a motivation for the study.

**MAP**

This research draws on a professional development program called the Management Achievement Program (MAP). MAP was not developed for the purpose of gathering data for analyzing retention. Instead, it is a program designed to develop and enhance the professional demeanor of its participants. The rationale for introducing MAP was that professionalism can best be developed and demonstrated when participants engage in co-curricular activities, in contrast to the assimilation of content delivered only through lectures or seminars. The philosophy of MAP is based on Experiential Learning and its premise that learning is “a cycle driven by the resolution of the dual dialectics action/reflection and experience/abstracts” [5], a philosophy embedded in how the program is defined:

> “An engaging and comprehensive program designed to develop and enhance each student’s professional demeanor, build competencies for academic success, increase involvement in the College and local business communities, and allow the opportunity for students to personally synthesize their academic and professional goals and experiences.”

MAP is a requirement for all undergraduate students in business administration at a state university. Students in MAP are required to select and participate in events and activities designed to build professional and career skills. Among the types of events and activities regularly offered are career workshops, seminars, forums, company visits, presentations by senior executives, student clubs, and service learning activities such as volunteer work.

Upon admission into the college, each participant’s record is assessed to determine the number of MAP miles to be required for completing the program. Miles are used as the metaphor for MAP participation, and once the MAP miles requirement has been established, a participant can start to earn miles to fulfill that requirement by participating in MAP events. Each event offers a number of miles depending on the involvement and initiative it requires. For the majority of events participants can earn 50 miles; an example of such an event would be a workshop on interviewing. Events such as competing in a college-wide case competition can earn as many as 200 miles. Some events also require participants to record their reflections to summarize what
they learned and how they envision applying it to their academic or professional careers. These written responses are also a source of valuable information that might be used in conjunction with the data in this study, but beyond the scope of this paper and part of further work.

The MAP program operates as follows. After matriculation and prior to their first semester, each participant is given a bar-coded identification card and access to a web-based portal for reviewing upcoming events, registering for specific events, entering reflections on events attended, and checking the status of MAP miles and accounts. Because MAP events are independent of course schedules and often held outside of classrooms or off-site, attendance and participation is recorded through a bar-coding scanning system which is regularly uploaded to a back-end system.

Figure 1 shows the infrastructure for supporting MAP; it should be noted that this figure depicts the operational component of MAP only.

Figure 1. Management Achievement Program (MAP) operational infrastructure

MAP was implemented in the fall of 2006, and during that initial semester, 40 events were offered and attended by 202 participations. To maximize the program’s benefits and offer participants flexibility, the number and type of events held each semester has grown significantly. During spring 2011, 2,466 participants were able to choose among 123 different events. MAP is now an integral part of the college’s curriculum and well-accepted in its culture; from inception through the end of August 2011, there have been 1,020 MAP events attended by a total of 17,317 participants.

Integrating professional and soft skills into a curriculum is held to be desirable but difficult to implement [10]. MAP has brought these directly into the college’s curriculum. Many institutions identify professionalism as a specific learning objective, but this is a learning outcome that is difficult to assess [7]. MAP has enabled professionalism and soft skills to be assessed with the use of methods such as Latent Semantic Analysis [2].

MAP has been successful as a program of co-curricular activities. But can MAP help predict student retention by analyzing how students participate in this program? That is purpose of this research, and the objective of the methodology discussed in the next section.
METHODOLOGY

Students are enrolled in MAP and made aware of their requirement for MAP miles and the operation of the program prior to beginning of their first semester with the college. The number of MAP miles they will need to earn to complete the program is based on a sliding scale, and is usually 250, 500, 750, or 1,000 miles. As noted earlier, students are then free to choose from the schedule of MAP events offered during and between semesters to complete their requirements, and the number of miles to be earned varies depending on the specific MAP event.

To evaluate whether or not participation in MAP might be related to retention and ultimately a successful academic career our first step was to determine if there exist groups whose progress through the program is similar, meaning to find whether distinct patterns of participation in MAP exist.

To find if there are any patterns in participation we conducted a cluster analysis. This analysis clustered participants using the percentage of their MAP miles requirement that was completed during each semester of their enrollment, starting from their first semester and extending through two years.

We ran numerous $k$-means cluster analyses and found that for values of $k$ higher than 3, the analysis generally produced one or more clusters having a small number of observations that were not viable. After examining these runs, we determined that the most meaningful results were produced by a value of 3 for $k$. Using that setting, an individual cluster analysis was conducted for participants who starting in the same semester and held the same requirement for MAP miles. For example, there were four cluster analyses conducted for students starting in spring 2008, one each for those with requirements of 250, 500, 750, and 1,000 MAP miles.

A sample of the results from the cluster analysis is shown in Figure 2, which shows the average percentage of miles completed for participants who started in fall 2006 with 500 miles as their requirement for miles.

![Figure 2: Clusters of MAP participation rates](image)

The patterns in participation rates shown in Figure 2 can be interpreted as follows. Each line represents a cluster of participants, each showing the average percentage of miles participants
accumulated towards their requirements during each semester. For example, participants represented by the solid line in the middle reached approximately 15% of their requirement during the first semester, 30% during their second, 60% during their third, and completed their requirement by reaching 100% during their fifth semester.

The patterns of participation seen in Figure 2 are similar to those found for participants starting in different semesters and having different requirements for MAP miles. What can generally be seen in the cluster analyses is that in one cluster participants are going well beyond the required miles, we tentatively named that cluster “High achievers.” The pattern for participation in another cluster seemed to be to fulfill just the essential requirements, and we termed that cluster “Meeting requirements.” The pattern for the third cluster was distinctly different from the other two, and a pattern in which participants fell well short of meeting their required miles. This cluster we termed “Retention risks.”

The apparently clear patterns that emerged from the cluster analysis, and the similarity of those patterns regardless of the semester a participant began their program or their requirements for MAP miles, led us to question whether the differences in participation rates for those clusters was statistically significant. To evaluate this we conducted one-way ANOVAs to test for significant differences between the participation rates in each cluster, i.e. between the percentages of miles completed over time for the participants in the three clusters.

The ANOVAs showed significant differences in participation rates; for example, for the 117 students starting in fall 2006 with a requirement of 500 miles, $p < .05$ as shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Clusters</td>
<td>.701</td>
<td>2</td>
<td>.351</td>
<td>12.056</td>
<td>.000</td>
</tr>
<tr>
<td>Within Clusters</td>
<td>3.315</td>
<td>114</td>
<td>.029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4.016</td>
<td>116</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Test of mean participation rate differences, participants with a MAP miles requirement of 500 and starting in fall 2006

The three clusters visually appear to represent distinctly different patterns of participation in plots, but to test whether those are statistically significant differences we preceded with planned post-hoc tests after significant differences were indicated by the omnibus ANOVAs. Table 2 presents the results of the post-hoc test for the same participants as analyzed for Table 1, students who started in fall 2006 with a requirement of 500 MAP miles. It confirms the impression of distinct differences seen in plots, finding significant differences between each of the three clusters of MAP participation, $p < .05$.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>N</th>
<th>Subset for alpha = 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 (&quot;Retention risks&quot;)</td>
<td>40</td>
<td>.0238</td>
</tr>
<tr>
<td>2 (&quot;Meeting requirements&quot;)</td>
<td>60</td>
<td>.1242</td>
</tr>
<tr>
<td>1 (&quot;High achievers&quot;)</td>
<td>17</td>
<td>.2618</td>
</tr>
</tbody>
</table>

Table 2: Results of Tukey-B test for homogeneous subsets of clusters
DISCUSSION

Of the three clusters from the results, the cluster “Meeting requirements” had the largest number of occurrences with slightly more than half of all participants. In general, these participants completed their miles requirements within four or five semesters. While we could not have expected this as a cluster, it is a logical result and consistent with participants generally completing the program. A finding of interest is that rather than simply fulfilling the requirement and ceasing participation, many participants in this cluster continued attending events and accumulating miles beyond the minimum required.

Participans in the “High achievers” cluster did not just meet their requirements for the program, but went far beyond them. During their eighth semester, on average they had exceeded requirements by an average of almost 50%. That this group would emerge as a cluster was not surprising, but the magnitude of the margin above the requirements and the size of the group - approximately 15% - were.

Some one-third of participants were in the third cluster which was termed “Retention risks.” These are students who, unlike those in the other two clusters, have not completed their requirements for the MAP program after two years, and with an average of half of their required miles completed, do not appear close to be completing them. While some students in this cluster may have completed more than half of their requirements, and some may ultimately complete the program, most do not. There are many different paths that students can take which can make categorization for retention difficult; some may switch from full-time to part-time, withdraw mid-semester from their courses, or stop taking courses for a semester or two only to return later. Regardless, the participants in this cluster clearly stand out as potential concerns for retention.

Finally, the analysis and the separation of participation rates evident in Figure 1 during the first semester suggest that whether a student will ultimately be in either the “High achievers”, “Meets requirements”, or “Retention risks” cluster may be apparent very early in an academic career, even during their first semester. Retention during the first semester is especially critical, yet many studies rely on factors only available after it has ended, such as GPA. While it is outside the scope of this paper, we are currently investigating this as further research among other directions as discussed next.

SUMMARY

The importance of student retention is widely recognized by colleges and universities and the impact of student attrition widely felt. Researchers have increasingly turned to analytical methods and data mining in particular to explain and predict student retention. The factors impacting retention that these studies have explored include demographics, family backgrounds such as the educational levels of family members, performance indicators such as GPA and scores on standardized tests, and amount of financial aid have all be used as factors to analyze retention.

Our research investigates how predictive data mining models for retention can be developed based on a program of co-curricular activities named MAP. Prior research of data mining models generally uses factors such demographics, education levels of family members, performance
measures, amount of financial aid to predict retention. When student engagement or involvement is studied as a factor, it is most often represented by the degree to which student services such as academic advising or tutoring are utilized. To our knowledge, this is the first study to examine how an institutional program of co-curricular activities might be used to measure and predict student retention. This program was designed for the purpose of developing professionalism and soft skills, which can be difficult to integrate explicitly into curricula and very difficult to assess. Our research shows that there are distinctly different patterns in participation rates, and using cluster analysis finds that participants tend to be in one of three groups, which we termed “High achievers”, “Meets expectations”, and “Retention risks.” Our research demonstrates that there is a statistically significant relationship between these participation patterns and retention.

This research is currently being extended in several significant directions. One of these is to build on the analytical methods presented in this paper. More specifically, we are augmenting the cluster analysis with other unsupervised techniques, such as association rules in a method from a seemingly unrelated but potentially applicable method developed in a study of ant colonization [6]. Other potential factors that might influence retention, including measures of performance such as GPA, are also being explored using supervised techniques to predict retention directly. Text analytics are the third category of techniques under development, which are being used to analyze the text of responses written by participants and the results integrated with those from the other techniques.

The statistically significant relationships found in our research raise the possibility that from participation in MAP, retention could be predicted very early in student’s career. This would have significant implications because attrition rates tend to be highest during the first year and identifying students at highest risk for early intervention is thought to be one of the most effective means to prevent attrition.

Finally, we are currently pursuing the relationship our research to existing theories. Tinto [12] explores several theories of student engagement that can be tested by our research, and theories from other fields are also very relevant. In particular, customer retention is equally as important to businesses and student retention is to universities, and businesses have found methods of engaging customers to be highly productive. The problem of retention and customer “churn” have many similarities [3], and some of the existing theories used to explain and predict customer retention are theories that might be tested or adapted to see how they apply to retention models based on student participation as presented in this research.

REFERENCES


